

---

# A Topic Model Approach to Multi-Modal Similarity

---

**Rasmus Troelsgård, Bjørn Sand Jensen and Lars Kai Hansen**

Department of Applied Mathematics and Computer Science

Technical University of Denmark

Matematiktorvet 303B, 2800 Kgs. Lyngby

{rast,bjje,lkai}@dtu.dk

## Abstract

Calculating similarities between objects defined by many heterogeneous data modalities is an important challenge in many multimedia applications. We use a multi-modal topic model as a basis for defining such a similarity between objects. We propose to compare the resulting similarities from different model realizations using the non-parametric Mantel test. The approach is evaluated on a music dataset.

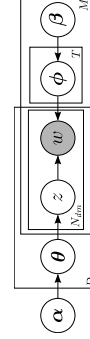
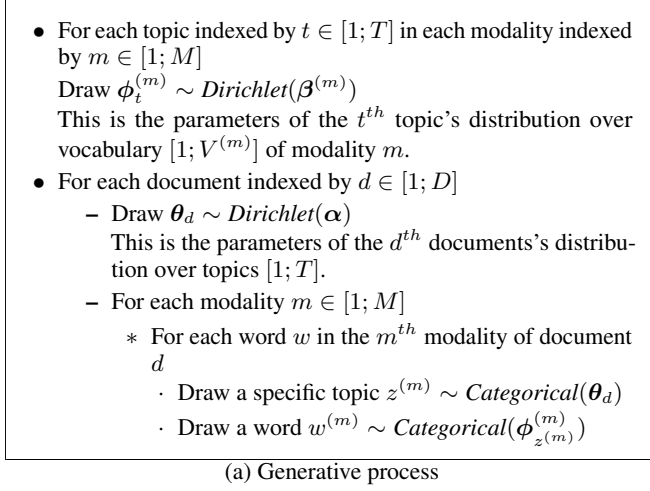
## 1 Introduction

Calculating similarity between objects linked to multiple data sources is more urgent than ever. A prime example is the typical multimedia application of music services where users face a virtually infinite pool of songs to choose from. Here choices are based on many different information sources including the audio/sound, meta-data like genre, and social influences [1], hence, attempts of modeling the geometry of music navigation have taken on a multi-modal perspective. In fusing heterogeneous modalities like audio, genre, and user generated tags it is both a challenge to establish a combined model in a 'symmetric' manner so that one modality do not dominate others and it is challenging to evaluate the quality of the resulting geometric representation. Here, we focus on the latter issue by testing the consistency of derived inter-song (dis-)similarity by means of direct comparison between similarities using the Mantel permutation test.

Topic models have previously been used to infer geometry in the image and music domain, e.g. by [2] combining audio features and listening histories. In [3] images and tags were analyzed, also by means of a multi-modal topic model. In [4] music similarity is inferred with a nonparametric Bayesian model, and [5] describe multiple multi-modal extensions to basic LDA models and evaluate the models on an image information retrieval task. Furthermore, topic model induced similarities among documents have been put to use in a navigation application [6], and different similarity estimates are also discussed in relation to a content-based image retrieval problem [7].

## 2 Model & Inference

To be able to measure similarities between objects, a representation of these objects is needed. In this work we use a version of Latent Dirichlet Allocation that incorporates multiple sources of information into a joint object representation similar to [5]. In [8], this model was applied to a multilingual corpus. Each object is represented by a multinomial distribution over topics which is common for all of the modalities composing the object. Each topic is defined by a set of multinomial distributions over features, each of which is defined on the vocabulary specific for a modality. To explain the characteristics of the model, the assumed generative process for objects is outlined in figure 1 together with a graphical representation of the model. The difference from a number of individual LDA models, each defined on a separate modality, is that each object is described by a single, shared distribution over topics, which potentially induces strong dependencies between the feature distributions representing the same topic in the individual modalities.



(b) The multi-modal Latent Dirichlet Allocation model represented as a probabilistic graphical model.

Figure 1

Performing inference in the model amounts to estimation of the posterior distributions over the latent variables. We use a Gibbs sampler inspired by the sparsity improvements proposed by [9]. For evaluation (see section 4), we use point estimates  $\theta^s$  and  $\phi^s$  derived from a sample  $\mathbf{z}^s$  from the Markov chain, by taking the expectations of the respective posterior Dirichlet distributions defined by  $\mathbf{z}^s$ . In this work we choose the state of the chain with the highest model evidence within the last 50 out of 4000 iterations. Hyper-parameters are optimized using fixed point updates [10, 11]. The prior on the document topic distributions is an asymmetric Dirichlet with parameter  $\alpha$ , and the priors over the vocabularies of the respective modalities are symmetric Dirichlet distributions with parameters  $\beta^{(m)}$ .

### 3 Similarities in Topic Models

As already hinted, there are many ways to define and calculate similarities in topic models; both between topics and documents. In this paper we focus on the latter. Most methods in literature are based solely on the distributions of topics in the documents,  $\theta$ , e.g. [4] measures the Kullback-Leibler divergence between two such distributions, while [7] also mentions inner products and cosine similarities as candidates. With focus on visualization, [6], introduces the yet another dissimilarity measure based on topic proportions. [7] promotes a measure based on the predictive likelihood of the document contents, and this approach is the basis of the method chosen here; The similarity of two documents  $A$  and  $B$  is given by the mean per-word log-likelihood of the words of document  $A$  given the topic distribution of document  $B$  (and the vocabulary distributions).

$$\frac{\log p(\mathbf{w}_A | \theta_B^s, \phi^s)}{\sum_{m=1}^M N_A^{(m)}}, \quad \text{where} \quad p(\mathbf{w}_A | \theta_B^s, \phi^s) = \prod_{m=1}^M \prod_{i=1}^{N_A^{(m)}} \sum_{t=1}^T (\phi_{t, w_{Ai}^{(m)}}^{(m)})^\top \theta_{t, B} \quad (1)$$

We use this approach to calculate a non-symmetric similarity matrix between all objects in the held-out cross-validation fold, for which the topic proportions have been estimated using “fold-in”.

<sup>1</sup> While this similarity measure is more computationally demanding than e.g. the KL-divergence, when the number of topics  $T$  used in the model increases, it might happen that some topics have vocabulary distributions that are very alike and only differ on a few words. Thus two documents with mainly the same type of content may have large proportions of different topics, causing them to be very dissimilar according to a topic proportion based measure. For a non-parametric topic model such as [4], this might not be a large concern, however, for parametric topic models, this should be taken into consideration. Generally, most of the discussed similarity measures are not proper metrics

<sup>1</sup>For the few held-out documents that do not contain any words in the modalities used for model estimation, we chose to simulate a uniform distribution of words in such an empty document by one occurrence of every word in the vocabulary.

in the geometric sense, but for (dis-)similarity purposes the exact properties might not be important, depending on the application.

#### Comparing Similarities - the Mantel test

An important aspect of this work is the ability to assess the relations between different similarities induced by models estimated from multiple, possibly different, heterogeneous data sources. To compare such similarities we look at the correlation between the defined similarities. For testing the significance of the correlations we can apply a Mantel style test [12]. The Mantel test is a non-parametric test to assess the relation between two (dis-)similarity matrices. The null hypothesis is that the two matrices are unrelated, and the null distribution is approximated by calculating the test statistic for a large number of random permutations of the two matrices (excluding the diagonal elements); permuting rows and columns together to maintain the distribution of (dis-)similarities for each object. In this work we use Spearman’s correlation coefficient as the test statistic.

### 4 Experimental Results: Music Similarity

In this preliminary study we examine induced similarities in a subset of the Million Song Dataset [13], consisting of 30.000 tracks with equal proportions of 15 different genres. Each track is composed of data from a number of different sources: Open vocabulary tags from users (last.fm), Lyrics (musiXmatch.com), Editorial artist tags (allmusic.com), Artist tags (musicBrainz), User listening history (echonest), Genre and style (allmusic), and Audio Features (echonest). All modalities—besides the audio features—are naturally occurring as counts of words and for the audio we turn to an *audio word* approach, where the continuous features are vector quantized into a total of 2144 words. For this pilot study we estimate topic models on combinations of groups of modalities from the mentioned list, respectively consisting of the first 5, the genre and style labels, and the audio. To be able to assess the model stability of the similarities, we estimate each model five times from different random initialisations of the Markov chain. This is done for every training set of a 10-fold cross-validation split. The correlations between all combinations of the 5 similarity matrices resulting from each held-out fold are then calculated, and the resulting distributions of correlation coefficients are shown in figure 2a. Figure 3a shows the distributions of correlations between similarities based on audio and on the larger modality group. The correlations are evidently much smaller than for identical models, but a Mantel test with 100 permutations suggest that the null hypothesis of no correlation can be rejected at a significance level of at least 1% for all three model complexities.

### 5 Discussion & Conclusion

The issue of stability is relevant for similarities induced by topic models using approximate inference techniques. The correlations between similarities from identical but randomly initialized models,

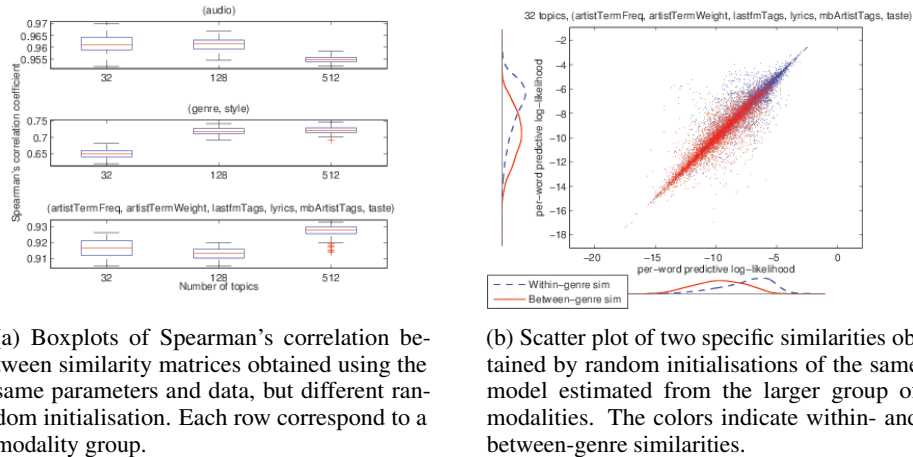
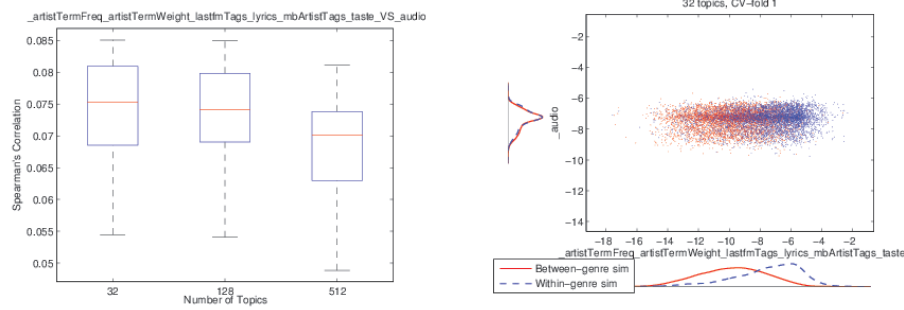


Figure 2



(a) Boxplots of Spearman's correlation between similarity matrices obtained using the larger modality group and the audio.

(b) Scatter plot of the two examples of similarities obtained from topic models with same parameters, but two mutually exclusive modalities of data.

Figure 3

can be used as a tool to gain some insight into this matter. From the preliminary results on the music example we find the induced similarities (fig. 2) to be highly stable. Furthermore, inspecting the similarities obtained from different data types; figure 3, we observe that while the audio model in itself does not seem to provide higher intra- than inter-genre similarity, it is still significantly positively correlated to the other modality group which does possess some discriminative power in terms of genre labels. Moreover, it seems that an increasing number of topics causes the correlation between similarities from models estimated on different modality groups to decrease. We speculate that this is linked to the specific topic model variant, for which [5] also note that the model describes the joint distribution of different modalities well, but does not model the relations between them.

In conclusion, we have proposed the multi-modal LDA as a method to define similarities in multimedia applications with multiple heterogeneous data sources based on the predictive-likelihood. This was extended with the Mantel test allowing direct evaluation of the consistency and correspondence of the resulting similarities.

### Acknowledgment

This work was supported in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328. This publication only reflects the authors' views.

### References

- [1] M.J. Salganik, P. Sheridan Dodds, and D.J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006.
- [2] Kazuyoshi Yoshii, M. Goto, K. Komatani, R. Ogata, and H.G. Okuno. Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*, pages 296–301, 2006.
- [3] Rainer Lienhart, Stefan Romberg, and Eva Hörster. Multilayer pLSA for multimodal image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 9:1–9:8, New York, New York, USA, 2009. ACM Press.
- [4] M. Hoffman, D. Blei, and P Cook. Content-based musical similarity computation using the hierarchical dirichlet process. *ISMIR 2008 - 9th International Conference on Music Information Retrieval*, pages 349–354, 2008.
- [5] D.M. Blei and M.I. Jordan. Modeling annotated data. *Annual ACM Conference on Research and Development in Information Retrieval*, pages 127–134, 2003.
- [6] A.J.B. Chaney and D.M. Blei. Visualizing Topic Models. *International AAAI Conference on Social Media and Weblogs*, 2012.
- [7] E. Hörster, R. Lienhart, and M. Slaney. Image retrieval on large-scale image databases. In *Proceedings of the 6th ACM international conference on Image and video retrieval - CIVR '07*, pages 17–24, New York, New York, USA, 2007. ACM Press.

- [8] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 2, 2009.
- [9] L. Yao, D. Mimno, and A. McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 937–946, New York, NY, USA, 2009. ACM.
- [10] T. P. Minka. Estimating a Dirichlet distribution. *Annals of Physics*, 2000(8):1–14, 2012.
- [11] H. M Wallach. *Structured Topic Models for Language*. PhD thesis, University of Cambridge, 2008.
- [12] N. Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2):209–220, 1967.
- [13] T. Bertin-Mahieux, D. P.W. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, 2011.